Supporting Information

da Cunha et al. 10.1073/pnas.0907939106

SI Methods

Public Data. The human genome sequence (National Center for Biotechnology Information build 36.1) was downloaded from University of California Santa Cruz Genome Bioinformatics portal (http://genome.ucsc.edu). A total of 26,281 sequences were downloaded from the Reference Sequence database (release 22; www.ncbi.nlm.nih.gov/RefSeq/). SNPs (5,689,286) were downloaded from dbSNP (build 127; www.ncbi.nlm.nih.gov/SNP/). A total of 372 SAGE libraries were downloaded from SAGE Genie (http://cgap.nci.nih.gov/SAGE). Thirty-seven MPSS libraries were downloaded from LICR-MPSS database (http://mpss.licr.org). Gene pathways were downloaded from KEGG (www.genome.jp/kegg/). Gene ontology data were downloaded from the GO database (www.geneontology.org).

Selection of Genes Coding for TM Domains. Two strategies were used to select known genes coding for TM domains. First, all translated Reference Sequence entries were submitted to Pfam (release 19.0) by using the program hmmm (1), and those sequences presenting domains annotated as TM were selected. Second, all translated Reference Sequence entries were submitted to the TMHMM web server (www.cbs.dtu.dk/services/TMHMM; default parameters), and those sequences presenting at least one TM domain were selected. TMHMM is the state-of-the-art algorithm for prediction of TM domains, with specificity and sensitivity levels above 99%. Pfam and TMHMM results were merged, and sequences containing a single domain in the first 50 amino acids were excluded (proteins classified as secreted). We also excluded sequences annotated as having a GO subcellular localization as cytoskeleton, endosome, golgi, liposome, lysosome, mitochondrion, nucleus, peroxisome, reticulum, or ribosome (and without a GO:0005886 for plasma membrane). See Fig. 1 for a schematic view of the strategy used in this report to select the surfaceome genes.

Mapping of Transcripts onto the Human Genome. Mapping of cDNA sequences onto the genome was performed by using the protocol described previously (2). In brief, all Reference Sequence entries were first mapped onto the genome sequence by using BLAT (3). Second, to improve the quality and specificity of the alignment, the best hit for each sequence in the genome was selected, followed by a pairwise alignment using Sim4 (4). Third, transcripts presenting identity greater than 95% and coverage greater than 90% were uploaded in a relational database.

Classification of Surfaceome Genes in GPCRs, SLCs, or CDs. To classify the surfaceome genes as GPCRs, SLCs, or CDs, we first selected all known members from these three gene categories. GPCR genes were retrieved from Gene Entrez (5), Gene Ontology database (GO:0007186, G-protein-coupled receptor protein) (6) and iuphar-db database (www.iuphar-db.org/). SLC genes were exclusively retrieved from Gene Entrez (5). CD genes were retrieved from the Human Cell Differentiation Molecules web site (http://hcdm.org/). After selection, genes from these datasets were compared to the set of surfaceome genes.

Gene Expression by MPSS and SAGE. To characterize gene expression by SAGE and MPSS, we first obtained a reliable tag for each gene (the methods and criteria for the selection of these tags were described previously in refs. 7 and 8). Then, this set of virtual tags was matched against experimental tags from 372 SAGE libraries and 37 MPSS libraries. Gene expression was determined by counting the tag frequency for each gene in the SAGE and MPSS libraries. SAGE libraries were normalized by 200,000 tags. MPSS libraries were normalized by 1 million tags, and only tags with a frequency greater than three were used.

Mutation in Colorectal and GBM Samples. Somatic mutation data were retrieved from Rand et al. (9), Sjoblom et al. (10), Wood et al. (11), and Parsons et al. (12), and their genomic coordinates were determined through a local mapping using BLAT. Genomic coordinates of mutations were crossed against the genomic coordinates of surfaceome genes, and each mutation was classified as intronic (located in the introns), coding (located in the coding region), or no coding (located in the untranslated regions). Nucleotides and amino acid substitutions reported by the mutations were also stored in our surfaceome knowledgebase.

Identification of the Surfaceome Set in Other Eukaryotic Species. Surfaceome genes conserved among 10 species were determined through the comparison of all surfaceome genes for each species and their ortholog genes. Surfaceome sets for *P. troglodytes*, *M. musculus*, *C. familiaris*, *B. taurus*, *G gallus*, *D. rerio*, *D melanogaster*, *C. elegans*, and *S. cerevisiae* were defined by using the same strategy applied to the identification of the human surfaceome set (see *Selection of Genes Coding for TM Domains*). Ortholog genes were determined by using the HomoloGene database (release 62; www.ncbi.nih.nlm.org/HomoloGene).

Hierarchical Clustering Analysis. Hierarchical clustering was performed by using Euclidian distance as the method for measuring distance and average linkage as the method for agglomerative measure. The heatmap package from R (www.r-project.org) was used to perform both calculations.

Mapping of Synonymous and Nonsynonymous SNPs. All SNPs available in the dbSNP (release 128) were mapped onto the human genome according to a strategy described previously by us (13). The classification of SNPs as synonymous or nonsynonymous was done based on the comapping of known Reference Sequence entries with annotated coding region. A PERL script was implemented to identify SNPs as synonymous or nonsynonymous and calculate their density in all different categories of genes.

Selection of Genes for qPCR Experimental Validation. The qPCR analysis was performed with two sets of genes: (*i*) genes showing a restricted expression in normal tissues and expression in at least one tumor, and (*ii*) genes showing differential expression in some tumors, including colorectal and GBM. Gene expression was calculated based on SAGE, MPSS, or cDNA data publicly available.

Samples. Eleven samples from colorectal tumors, seven samples from matched normal colon adjacent to the tumor resection, 11 samples from glioblastoma tumors, and 15 samples from non-tumoral brain were obtained from patients treated at Hospital Oswaldo Cruz and Hospital das Clínicas after explicit informed consent and with local ethics committee approval. Human glioma (t98g, a172, hog, and u87mg) and colorectal tumor (hct116, hct8, ls147, sw480, dld1, widr, colohsr320, colodm320, and ls180) cell lines were obtained from American Type Culture Collection and cultured following fabricant instructions. Details on all of the samples are shown in Table S3. Total RNA derived from 21 normal human tissues (thymus, prostate, fetal brain, trachea, skeletal muscle, fetal liver, uterus, small intestine, heart, bone marrow, kidney, stomach, liver, spleen, spinal cord, lung, testis, placenta, whole brain, breast, and colon) was purchased from Clontech.

Primer Design. The sequences of 902 RefSeqs comprising the selected surface genes and the reference genes rpl27 (NM_000988.3; ribosomal protein l27) and bcr (NM_021574.2; breakpoint cluster region) were used to design primers and to select a specific Universal Probe Library (UPL) probe (Roche) by using the Universal ProbeLibrary Assay Design Center (http://qpcr.probefinder.com/organism.jsp). Primers were designed by automatically selecting an intron-spanning assay. All primers and probes used are shown in Table S2.

RNA Extraction, cDNA Synthesis, and qPCR. Total RNA was prepared from cultured cells and from tissues by using TRIzol (Invitrogen). RNA samples were checked for integrity by Bioanalyser (Agilent). The cDNA synthesis was carried out with 200 ng of DNA-free RNA by using SuperScript III First-Strand Synthesis SuperMix (Invitrogen). A preamplification reaction was performed by pooling all primers (final concentration, 50 nM), 5 μL of cDNA, and 10 μL of 2× PreAmp Master Mix (Applied Biosystems). The cycling program consisted of 95 °C for 10 min, followed by 14 cycles of 95 °C for 15 sec and 60 °C for 4 min. At the end, the reactions were diluted 1:5 with TE buffer (10 mM Tris-HCl pH 8 and 1 mM EDTA). The qPCRs were performed in 96.96 dynamic array chips (Fluidigm) following the fabricant instructions. Briefly, for each individual assay, a 10× Assay Mix that contains 2 μM forward primer, 2 μM reverse primer, 1 μM UPL probe, and 0.025% Tween-20 was prepared, and 5 μL of the mix was loaded into one of the Assay Inlets of the M96 Dynamic Array. In the sample inlets, 5 μL of the following solution was dispensed: 2.5 μL of PreAmp sample, 3.25 μL of 2× AB Universal TaqMan Master Mix (Applied Biosystems), and 0.32 μL of 20× DA Sample Loading Solution (Fluidigm). The cycling program used consisted of 2 min at 50 °C, 10 min at 95 °C, followed by 35 cycles of 95 °C for 15 sec, 70 °C for 5 sec, and 1 min at 60 °C. All reactions were performed four times. The CT values were obtained by using the BioMark Gene Expression Data Analysis after automatic inspection for quality. CTs higher than 30 and low-quality reactions were excluded and considered as not available. Relative gene expression values were determined by using the $2^{-\Delta\Delta CT}$ method of Livak and Schmittgen (14). Rpl27 and bcr were used as reference genes. Normal whole brain (from Clontech) was used as a reference sample for all normal glioma cell lines and GBM samples, whereas normal colon (from Clontech) was used as a reference sample for all normal colon cell lines and colorectal samples. Genes were considered overexpressed in GBM when three or more GBM samples (derived from patients) had a fold change at least three times higher than the standard deviation in at least two qPCR experiments. For the genes selected by using this criterion, we next averaged the three qPCR experiments on only those genes whose average fold was higher than three times the standard deviation. The same strategy was used for colorectal samples. Genes expressed in five or fewer normal samples were considered restrictedly expressed. To be considered expressed, the gene/sample must have had any valid CT in two qPCR experiments. Expressions in fetal brain, fetal liver, brain, placenta, and testis were not considered in this analysis. Genes expressed (in any valid CT) in testis and/or whole brain and/or placenta and in three or fewer normal samples (in at least two qPCR experiments) were considered putative CT, CB (cancer/brain), or CP (cancer/placenta), respectively. Expressions in fetal brain and fetal liver were not considered in this analysis.

1. Finn RD, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36:D281–D288.
2. Galante PA, et al. (2007) Sense-antisense pairs in mammals: Functional and evolutionary considerations. *Genome Biol* 8:R40.
3. Kent WJ (2002) BLAT–the BLAST-like alignment tool. *Genome Res* 12:656–664.
4. Florea L, et al. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* 8:967–974.
5. Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res* 35:D26–D31.
6. Ashburner M, et al. (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25–29.
7. Boon K, et al. (2002) An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci USA* 99:11287–11292.
8. Galante PA, et al. (2007) Automatic correspondence of tags and genes (ACTG): a tool for the analysis of SAGE, MPSS and SBS data. *Bioinformatics* 23:903–905.
9. Rand V, et al. (2005) Sequence survey of receptor tyrosine kinases reveals mutations in glioblastomas. *Proc Natl Acad Sci USA* 102:14344–14349.
10. Sjoblom T, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–274.
11. Wood LD, et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318:1108–1113.
12. Parsons DW, et al. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* 321:1807–1812.
13. Silva AP, et al. (2004) The impact of SNPs on the interpretation of SAGE and MPSS experimental data. *Nucleic Acids Res* 32:6104–6110.
14. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods* 25:402–408.

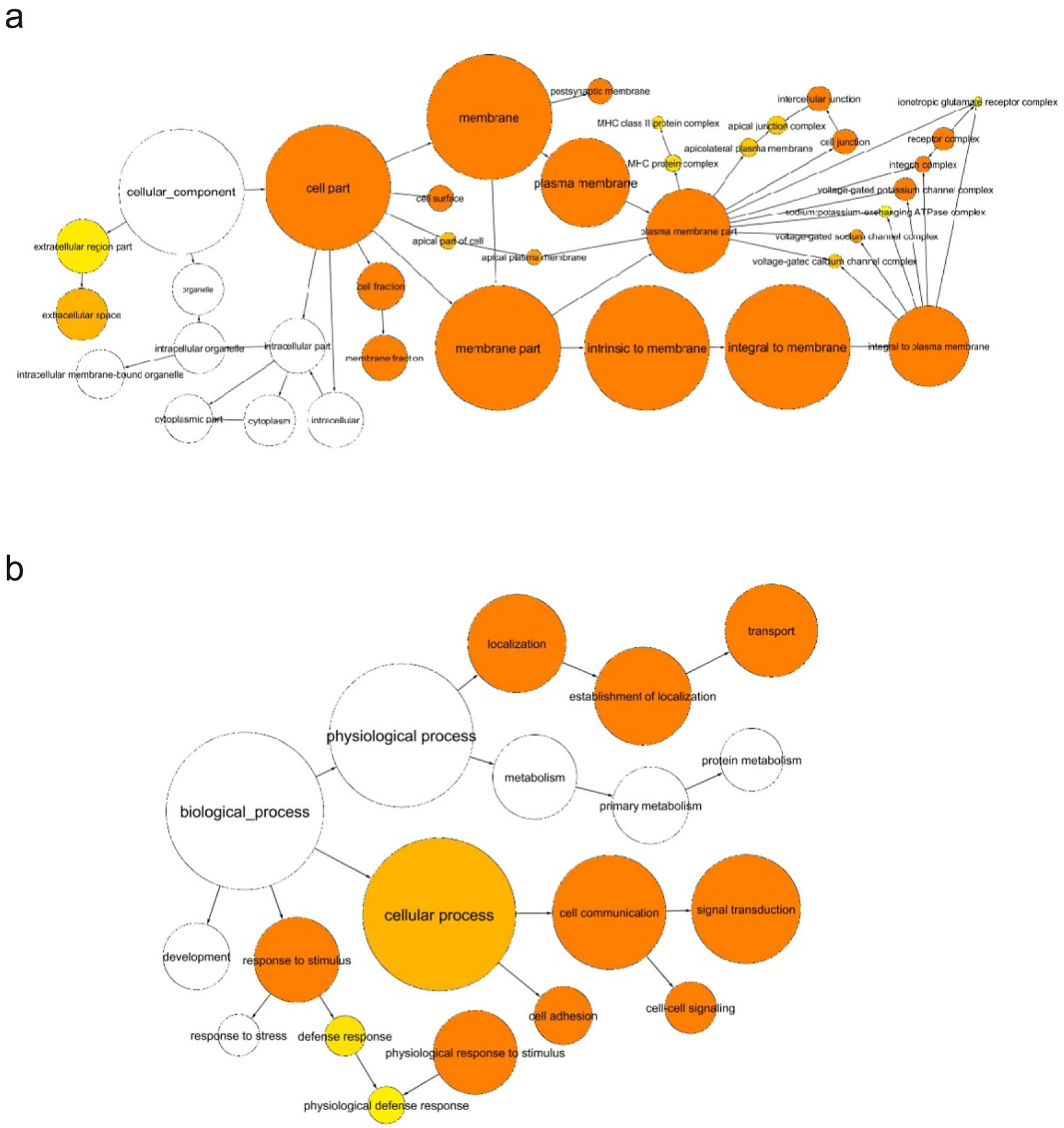

Fig. S1. GO classification of surfaceome genes in molecular function (*a*) and biological process categories (*b*). The whole collection of surfaceome (3,702) genes was classified by using BiNGO (www.psb.ugent.be/cbd/papers/BiNGO/). Diameter of the circle corresponds to the number of genes in each category. Color gradient corresponds to the significance of this enrichment (from white, not significant, to orange, the most significant).

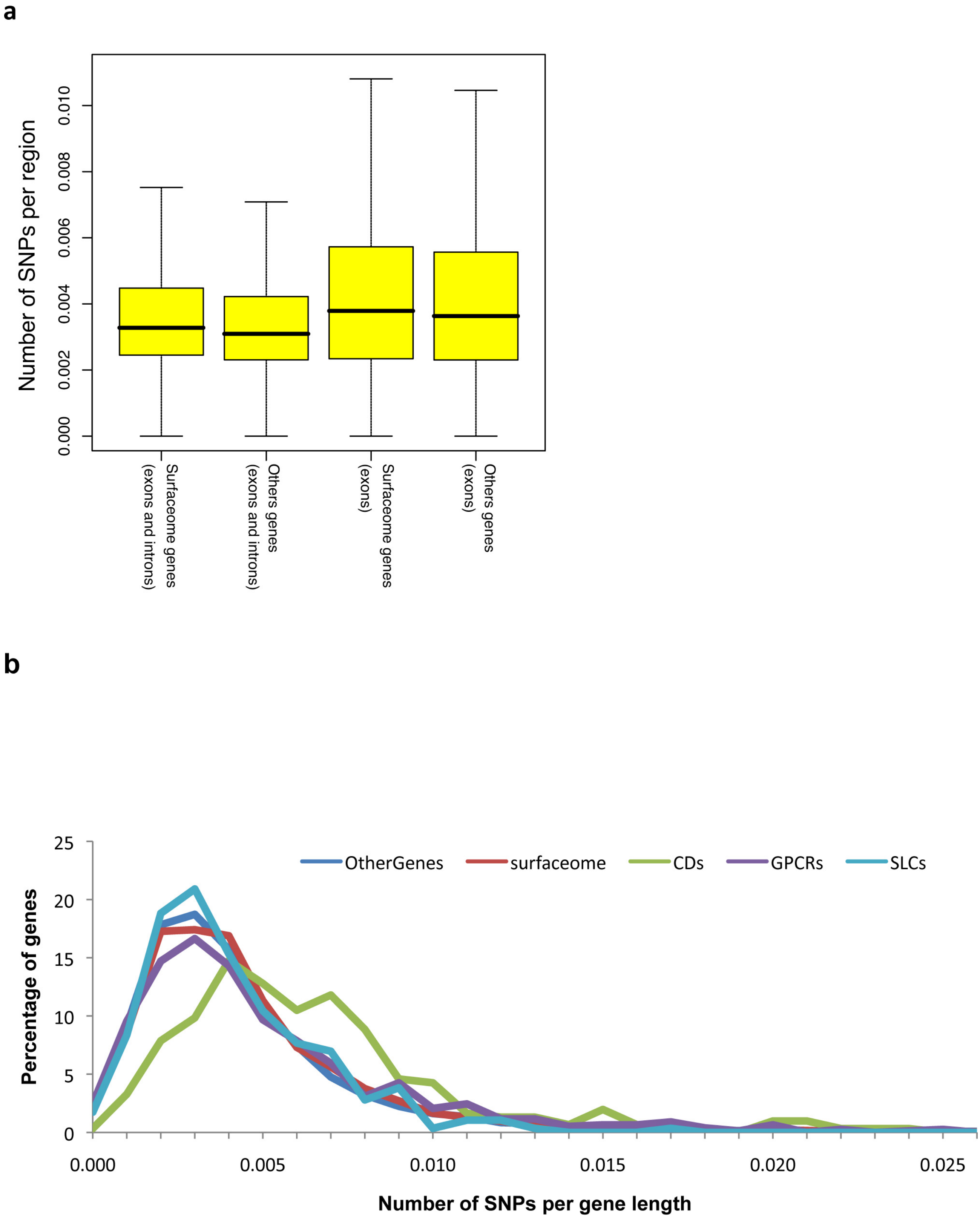

Fig. S2. SNP distribution in human genes. (*A*) Boxplot of distribution of SNPs in four datasets: surfaceome genes (exons only and exon plus intron) and remaining genes (exons only and exon plus intron). No significant difference was observed between surfaceome and remaining genes. (*B*) Coding SNP density distribution for gene categories.

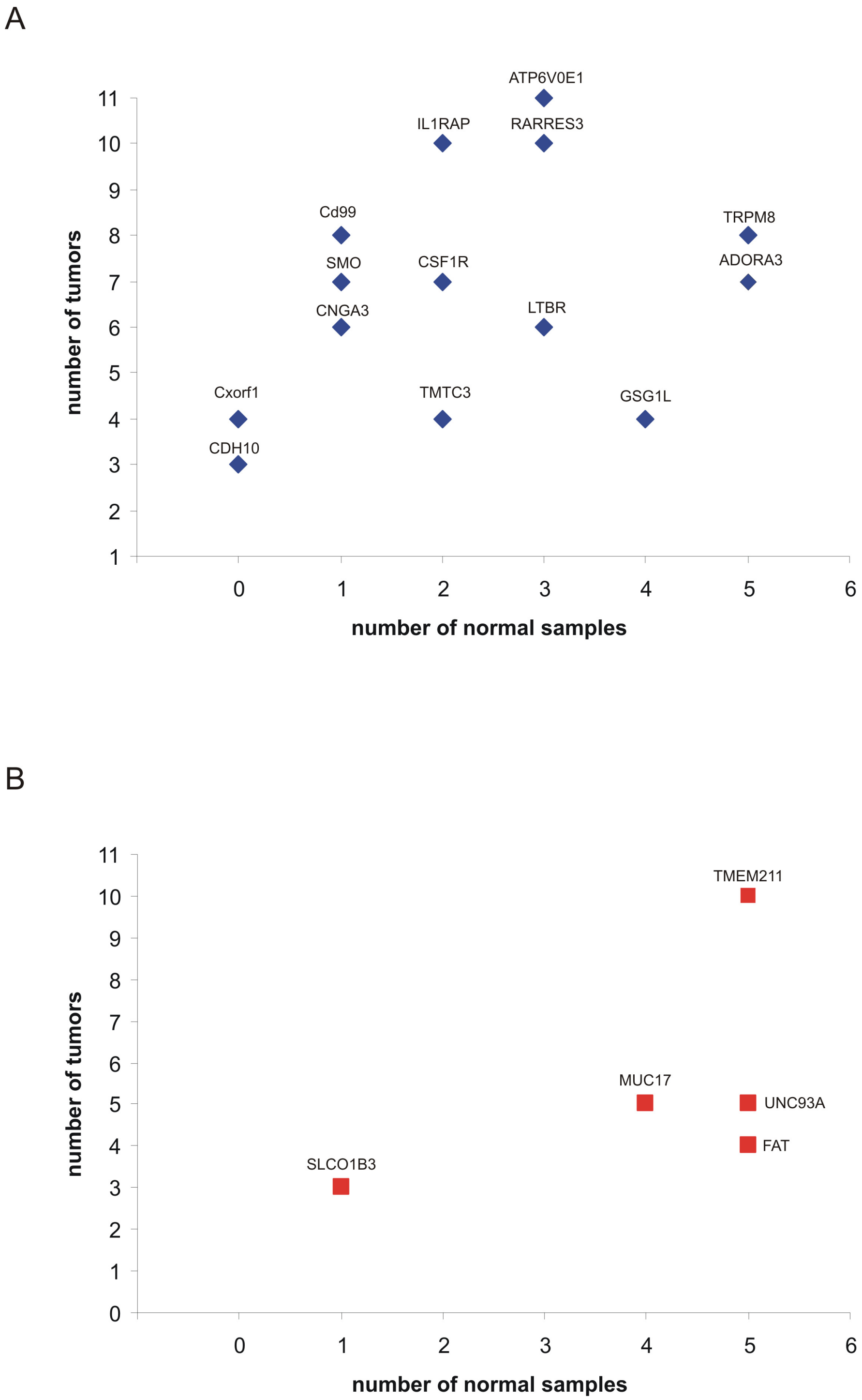

Fig. S3. Genes restrictedly expressed in normal tissues and overexpressed in GBM (*A*) and colorectal tumors (*B*). The *x* axis represents the number of normal tissues (excluding testis, fetal tissues, brain, and placenta) in which a given gene is expressed, whereas the *y* axis represents the number of tumor samples with fold values at least three times higher than the standard deviation. The number of normal tissues was inferred by any CT value in a given sample in at least two qPCR experiments.

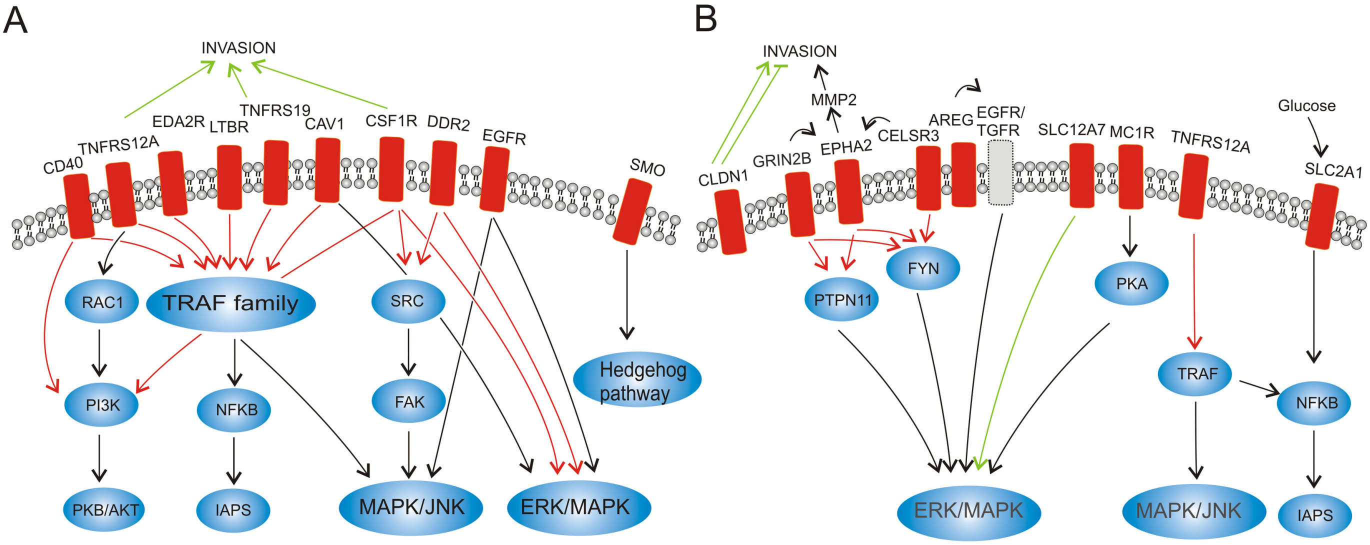

Fig. S4. Major surfaceome signaling pathways in GBMs (*A*) and colorectal tumors (*B*). Red arrows represent interactions derived from PPI networks. Black arrows represent interactions derived from KEGG, whereas green arrows represent interactions derived from searches in the literature.

Other Supporting Information Files

Table S1 (XLS)
Table S2 (PDF)
Table S3 (PDF)
Table S4 (PDF)
Table S5 (PDF)